# 聚类学习与总结

162010113 花国栋

**摘 要：**在进行课堂上进行模式识别相关算法的学习时，聚类算法是我为数不多能够听得比较明白的算法，而且这个算法并不要求我拥有一定的人工智能基础，在我看来就是处理数据的一种手段，直观且简洁明了，因此我对聚类算法产生了浓厚的兴趣。经过查找资料并结合现有知识，我了解了聚类的各种类别以及度量聚类结果的标准等知识，明白了各个聚类算法彼此的差别，在本文中我会整理出本人的学习成果以及本人理解的各个算法，或许并不准确，若有不对之处，还请老师指正。

**关键词：**聚类算法、相似度度量、欧氏距离、K-means、层次聚类

1 聚类(Clustering)的概念

聚类是将一堆数据对象按照特征相似性进行分组，使得组内的特征相似度相对较大，组间的特征相似性相对较小，这个组叫做簇(Cluster)。

聚类与分类的区别：

聚类：是无监督学习(Unsupervised Learning)，把相似的数据划分到一起

分类：是监督学习(Supervised Learning)，把不同的数据划分开，是通过已知分类的数据集进行训练，再通过学到的方法进行未知数据的预测分类。

2 聚类的过程

(1)数据准备：包括特征标准化和降维。

(2)特征选择：从最初的特征中选择有效的特征，存储到向量中。

(3)特征提取：通过对所选择的特征进行转换形成新的突出特征。

(4)聚类：基于某种距离函数进行相似性度量，然后进行聚类。

(5)聚类结果评估：对聚类的结果进行优劣的评估，如距离误差

3 聚类结果的度量

对于聚类结果的度量指标包括内部指标与外部指标。

3.1内部指标

不涉及任何外部信息的条件下，也不借助外部参考模型，仅依赖数据集自身的特征和度量值，通过计算簇内部平均相似度、簇间平均相似度或整体相似度来评价聚类结果的优劣和判断簇的最优个数。理想的聚类结果是簇内紧密且簇间分离，因此，常用内部指标的主要思想是通过簇内(intra-cluster)距离和簇间(inter-cluster)距离的某种形式的比值来度量的[1]。而度量两个n维样本和距离的公式主要有以下几种：

（1）欧氏距离(Euclidean Distance)：计算欧氏空间中两点的直线距离，计算公式如下



（2）曼哈顿距离(Manhattan Distance)：可以理解为在地图中从一个点到另一个点只能沿着横平竖直的街道的距离之和，计算公式如下



（3）切比雪夫距离(Chebyshev Distance)：是向量空间中的一种度量，二个点之间的距离定义是其各坐标数值差绝对值的最大值，计算公式如下



(4) 闵可夫斯基距离(Minkowski Distance)：是欧式空间的一种测度，是一组距离的定义，被看作是欧式距离和曼哈顿距离的一种推广，计算公式如下



是一个可变的参数。可以看出，时是曼哈顿距离，时是欧氏距离，时是切比雪夫距离。闵可夫斯基距离不是一种距离，而是一类距离。

基于以上几种距离度量，可以得到以下几种主要的聚类性能度量内部指标：

（1）紧密度(Compactness)：每个簇中的样本点到聚类中心的平均距离。



是簇大小，是这个簇的聚类中心。

一般使用所有簇紧密度的平均值来衡量聚类结果的好坏



为簇的总数。紧密度的值越小，表示簇内样本点的距离越近，即簇内样本的相似度越高,聚类结果越好。

（2）分隔度(Seperation)：是各簇的聚类中心两两之间的平均距离



分隔度的值越大，表示各聚类中心相互之间的距离越远，即簇间相似度越低，聚类结果越好。

（3）戴维森堡丁指数(Davies-Bouldin Index，简称DBI)[2]：计算任意两类别的类内距离平均距离之和除以两聚类中心距离求最大值。DB越小，意味着簇内距离越小同时簇间距离越大。





其中，,表示簇的每个点与该簇的聚类中心之间的平均距离，​表示聚类和的质心之间的距离。值越小，簇内相似度越高，簇间相似度越低，聚类结果越好。

（4）邓恩指数(Dunn Index，简称DI)[3]：任意两个簇的样本点的最短距离与任意簇内样本点的最大距离的比值。



为簇大小，与为不同聚类，值越大代表簇内相似度越高，簇间相似度越底，聚类效果越好。

3.2 外部指标

将聚类结果与某个参考模型进行比较，称为外部指标。对于含有个样本点的数据集，其中的两个不同样本点和假设C是聚类算法给出的簇划分结果，P是外部参考模型给出的簇划分结果。记为下面四个数[4]。

,：,在C和P中属于相同的簇。

,：,在C中属于相同的簇，在P中属于不同的簇。

,：,在C中属于不同的簇，在P中属于相同的簇。

,：,在C和P中属于不同的簇。

（1）Jaccard系数(Jaccard Coefficient，简称JC)



(2)FM指数(Fowlkes and Mallows Index，简称FMI)



（3）Rand指数(Rand Index,简称RI)



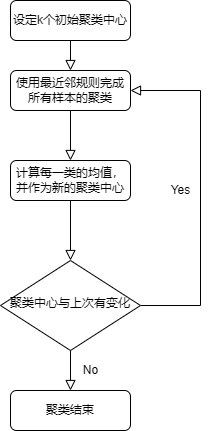
上述性能度量的结果都在区间，值越大代表聚类结果越好。

4 聚类算法

传统的聚类算法主要分为基于划分的聚类、基于层次的聚类、基于密度的聚类、基于网格的聚类、基于模型的聚类。[5]

4.1 基于划分的聚类算法

基于划分的聚类方法知识是将数据对象集合划为为若干个无交集的子集（簇），使得每个对象仅属于一个子集。主要包括K均值(K-means)算法、K-modes算法、围绕中心点划分(Partitioning Around Medoids,简称PAM)算法和大型应用中的聚类算法(Clustering LARge Applications,简称CLARA)等[6]。在这里主要谈谈K-means[7]算法。

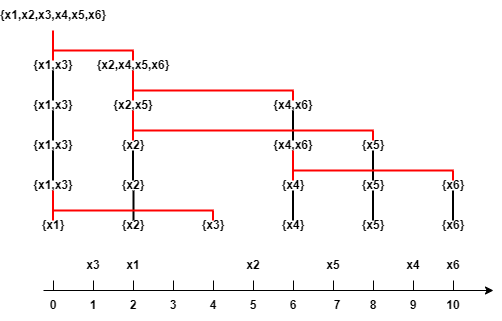


图表 1 K-means流程图

K-means算法步骤：

1. 选择初始化的个样本作为初始聚类中心。
2. 针对数据集中每个样本计算它到个聚类中心的距离并将其分到距离最小的聚类中心所对应的类中。
3. 针对每个类别，重新计算它的聚类中心（即属于该类的所有样本的质心）。
4. 重复第2、3两步操作，直到达到某个终止条件（例如迭代次数、最小误差变化等）。

4.2 基于层次的聚类算法

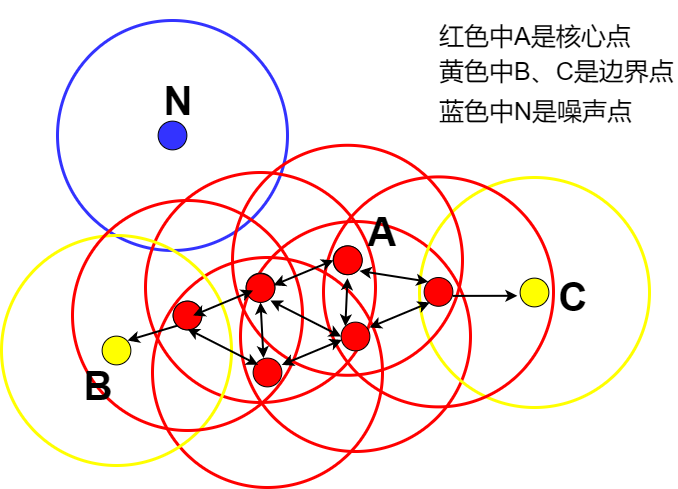
层次聚类的基本思想是构建数据之间的层次关系以进行聚类，所得到的聚类结果具有树形结构特征，称之为聚类树(Clustering Tree)。层次聚类分为两大类：自底向上(bottom-up)的层次聚类和自顶向下(top-down)的层次聚类。自底向上的层次聚类又叫聚合聚类，最初将每个数据点视为孤立簇（仅包含一个节点），然后根据簇之间的距离将最近的两个簇合并，直到只剩下一个簇；自顶向下的层次聚类又叫分裂聚类，它与聚合聚类相反，将整个数据集由大簇逐步拆分为小簇[8]。出于效率的考虑，目前大多数采用聚合聚类而非分裂聚类，主要包括AGNES(AGglomerative NESting)算法、BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies)算法、CURE(Clustering using Representative)算法等。在这里主要谈谈AGENES算法。

图表 2 聚合算法过程

AGENES算法步骤：

1. 将数据集中每个样本单独作为一个簇。
2. 找到簇间距离最小的两个簇，合并为一个簇，并重新计算这个簇的聚类中心。
3. 重复第2步操作，直到达到某个终止条件（例如聚类数量到达预设值、最近的簇间距离大于阈值）。

4.3基于密度的聚类算法

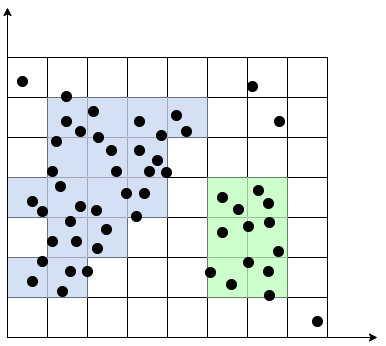
基于划分聚类和基于层次聚类的方法在聚类过程中根据距离来划分类簇因此只能够用于挖掘球状簇。基于密度的聚类算法不仅可以处理球状簇，还可以处理非凸型数据和发现任意形状的簇。这类方法形成的簇是被低密度区域分割的密集区域，利用密度函数把连续且密集点分为一类，将处于低密度区域的样本点标记为噪声[9]。主要算法包括DBSCAN(Density-Based Spatial Clustering of Applications with Noise)算法、OPTICS(Ordering points to identify the clustering structure)算法、DENCLUE(DENsity-based CLUstEring)算法。这里主要谈谈DBSCAN[10]算法。

图表 3 DBSCAN算法过程

DBSCAN算法步骤：

1. 以每一个样本点为圆心，以为半径画一个圆。这个圆覆盖的区域是的邻域。
2. 若的邻域内的样本点数量大于等于密度阈值，则记为核心点。若的邻域内的样本点数量没有超过密度阈值但是处于其他核心点的邻域内，则记为边界点。在样本集中既不是核心点也不是边界点的点就是噪声点。
3. 核心点对其邻域内样本点都是密度直达，若有对密度直达、对密度直达，则称对密度可达；由上面的假设，有与均由密度可达，则称和密度相连。将所有密度相连的点加入同一个簇，得到最后的聚类结果。
   1. **基于网格的聚类算法**

原理是将数据空间划分为网格单元，将数据对象映射到网络单元中，并计算每个单元的密度。根据预设阈值来判断每个网格单元是不是高密度单元，由邻近的稠密单元组成簇。由此可见，网格方法本质上只能看作一种压缩手段，它必须与密度结合起来才能进行聚类分析[11]。典型算法包括STING(Statistical Information Grid)算法、CLIQUE(Clustering In QUEst)算法、小波聚类（WaveCluster）算法等。



图表 4 基于网格的聚类算法过程

此类算法步骤：

1. 将数据空间划分为网格单元。
2. 使用网格单元内数据的统计信息对数据进行压缩表示。
3. 基于这些统计信息以及预先设定的阈值判断高密度网格单元。
4. 最后将相连的高密度网格单元识别为簇。
   1. **基于模型的聚类算法**

基于模型的聚类方法，主要是指基于概率模型和基于神经网络模型的方法。前者的代表是高斯混合模型(Gaussian Mixture Model,简称GMM)，后者的代表是自组织映射神经网络(Self-organizing map, 简称SOM)。这里主要谈谈高斯混合模型。通俗来说任何一个曲线，无论多么复杂，我们都可以用若干个高斯曲线来无限逼近它，这就是高斯混合模型的基本思想。

高斯混合模型算法步骤：

1. 初始化高斯混合成分的个数，假设高斯混合分布模型参数 (高斯混合系数)、（均值）、（协方差矩阵）
2. 分别计算每个样本点的后验概率（该样本点属于每一个高斯模型的概率）
3. 迭代,,
4. 重复第2步直到收敛[4]

5 结语

各种各样的聚类算法层出不穷，没有绝对的第一的算法，只能根据不同的领域与情景选择合适的算法进行使用。此外，明确每种算法的优缺点，有助于在将算法实际运用时明白如何对输入数据进行恰当的处理，而不会出现对数据各种降维方法使用一遍并且各种算法使用一遍这种穷举来找到最佳方法的情景，但这也并不是说实验是多余的，恰恰相反，实验得到的数据是必要的，它们是根据实际需求选择与改进算法的必要条件。

此外，对于算法的应用不可以一板一眼，并非说使用某种算法就一定是恪守这种算法的使用模板，像上面提到的基于网格的聚类算法，实际就是网格与密度结合的聚类算法。学习聚类算法，并不止是学习算法的使用，要学习算法的思想，学习处理数据的角度，这样才能在实际运用时融会贯通的使用算法。

参考文献

1. ZOU C S,DUAN G Q. Research on clustering quality evaluation index based on improved K-medoids[J]. Computer System Applications,2019,28(06):235-242.

邹臣嵩,段桂芹.基于改进K-medoids的聚类质量评价指标研究[J].计算机系统应用,2019,28(06):235-242.

1. Davies DL, Bouldin DW. A cluster separation measure.IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, 2(2): 224–227.
2. J. C. Dunn (1973) A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, Journal of Cybernetics, 3:3, 32-57
3. ZHOU Z H. Machine Learning[M].TsingHua University Press,2016.

周志华. 机器学习[M]. 清华大学出版社, 2016.

1. LIU W. A survey of clustering algorithms in data mining[J] Jiangsu Commercial Forum,2018,(07):120-125.

刘维.数据挖掘中聚类算法综述[J].江苏商论,2018,(07):120-125.

1. WANG Y H,LUO D S L. A survey of clustering algorithms[J] Science and technology information, 2018,16(24):10-11.

王玉晗,罗邓三郎.聚类算法综述[J].科技资讯,2018,16(24):10-11.

1. MacQueen J. Some methods for classification and analysis of multivariate observations[C], in: Neyman J, eds. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. United States :University of California Press, 1967. 281-297.
2. XIE W B.Research on hierarchical clustering algorithm based on reciprocal nearest neighbor and its application[D]Sichuan: University of Electronic Science and Technology of China,2021

谢文波.基于互惠最近邻的层次聚类算法及其应用研究[D].四川：电子科技大学,2021.

1. SONG P. Research and application of density-based clustering algorithm[D]Jiangsu:Jiangnan University,2022

宋鹏. 基于密度的聚类算法研究与应用[D].江苏：江南大学,2022.

1. Martin Ester, Hans-Peter Kriegel ,et al. A density-based algorithm for discovering clusters in large spatial databases with noise . [C]Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231.
2. WU Y H. A survey of clustering algorithms[J] Computer Science,2015,(S1):491-499+524.

伍育红.聚类算法综述[J].计算机科学,2015,(S1):491-499+524.